

# Biomedical text summarization to support genetic database curation: using Semantic MEDLINE to create a secondary database of genetic information

**T. Elizabeth Workman, MLIS; Marcelo Fiszman, MD, PhD; John F. Hurdle, MD, PhD; Thomas C. Rindflesch, PhD**

See end of article for authors' affiliations.

DOI: 10.3163/1536-5050.98.4.003

**Objective:** This paper examines the development and evaluation of an automatic summarization system in the domain of molecular genetics. The system is a potential component of an advanced biomedical information management application called Semantic MEDLINE and could assist librarians in developing secondary databases of genetic information extracted from the primary literature.

**Methods:** An existing summarization system was modified for identifying biomedical text relevant to the genetic etiology of disease. The summarization system was evaluated on the task of identifying data describing genes associated with bladder cancer in

MEDLINE citations. A gold standard was produced using records from Genetics Home Reference and Online Mendelian Inheritance in Man. Genes in text found by the system were compared to the gold standard. Recall, precision, and F-measure were calculated.

**Results:** The system achieved recall of 46%, and precision of 88% (F-measure=0.61) by taking Gene References into Function (GeneRIFs) into account.

**Conclusion:** The new summarization schema for genetic etiology has potential as a component in Semantic MEDLINE to support the work of data curators.

## INTRODUCTION

Due to evolving technologies and policies, libraries have an increasing interest in the process of data curation. As McDonald and Uribe point out [1], the open access movement, coupled with ever-increasing volumes of data from current scientific investigations, has created a research environment that calls for new management strategies for domain-specific data curation, which is defined here as the organization, preservation, and enhancement of the data through value-added features such as annotations. This environment has united traditional academic participants such as librarians, researchers, and administrators, who previously worked independently.

Librarians have the opportunity to take a leadership role in implementing techniques and policies for data curation and preservation. For example, an academic library could partner with other campus departments in creating the framework for enhancing and preserving the institution's research, possibly creating unique and priceless resources. There are several examples in which librarians have taken the lead in information curation, access, preservation, and management, including in neuro-ophthalmology [2], institutional repositories [3], and other areas. Curators of secondary databases face the demanding task of identifying relevant information from primary sources, which are continually increasing [4]. The development of curated databases is often based on a complex methodol-

### Highlights

- Semantic MEDLINE streamlines information retrieval by succinctly expressing the meaning of sometimes complicated text and summarizing output according to a user's needs.
- Semantic MEDLINE identifies genes noted in biomedical text as associated with a disease process.
- Semantic MEDLINE can potentially simplify secondary database curation.

### Implications

- Library information retrieval services can potentially benefit from automated applications such as Semantic MEDLINE.
- Use of such automated applications can facilitate the library's work in interdepartmental collaborative endeavors, thus reinforcing the library's core value in its parent institution.

ogy of information discovery, content development, and expert review [5, 6].

Information discovery for secondary databases may depend on traditional information retrieval and the meticulous, manual inspection of documents resulting from conventional searches of databases such as MEDLINE. This task can be quite daunting and time consuming. In developing the Human Protein Reference Database (HPRD), for example, developers performed extensive searches in PubMed to identify relevant literature. Then, researchers spent over



This article has been approved for the Medical Library Association's Independent Reading Program <<http://www.mlanet.org/education/irp/>>.

50,000 hours during an 8-month period reading more than 300,000 articles to manually curate HPRD records [7].

Biomedical information retrieval techniques provide support for secondary database curation [8]; however, little research has been published on using automatic summarization to augment these techniques and help manage the information contained in the large numbers of MEDLINE citations often returned by PubMed searches. Automatic summarization provides the information most relevant to a user's interest from a source in a condensed format. The advanced biomedical information management application Semantic MEDLINE\* [9] integrates automatic summarization with information retrieval, semantic processing, and visualization to analyze biomedical text. Semantic processing in the application uses SemRep [10, 11] to represent document content as semantic relations (e.g., drug X TREATS disease Y), also referred to as semantic predications. Automatic summarization [12] further processes these relations to identify those that are most relevant to a user's needs. The resulting semantic relations are then presented to the user in a graph that visually displays the content of retrieved documents. Because links are maintained between semantic relations and input text, the graph serves as a guide to help users decide what to read.

The thrust of the research reported here was to extend the use of Semantic MEDLINE to the domain of molecular genetics. Librarians maintaining databases in this domain must keep pace with the growing amounts of data generated by improved genetic analytic technologies [13] and need the ability to easily identify genes associated with a particular disease. The authors first describe the technology required to extend Semantic MEDLINE and then suggest how the application can serve as an adjunct to traditional information retrieval in secondary database curation. In the evaluation, genes extracted by the system were compared to those found in two actively curated genetic databases, Genetics Home Reference and Online Mendelian Inheritance in Man (OMIM).

## BACKGROUND

### Curated resources

Genetics Home Reference [14], hosted by the National Library of Medicine, was introduced in 2003 as a consumer-friendly website for genetic diseases [15]. The site implements a content development strategy that combines human effort with select complementary automated functions [16]. The OMIM database [17], a Johns Hopkins University product hosted by the National Center for Biotechnology Information at the National Library of Medicine, implements a

curation strategy in which journal content is reviewed daily by hand [18, 19]. Under agreement with publishers, OMIM receives articles from specific journals prior to publication. OMIM staff also read additional publications looking for potential materials for manual review. Genetics Home Reference provides information on a level appropriate for patients; OMIM furnishes more technical, detailed genetic disease information suited for scientists. The two databases provide a full landscape of online genetics information.

### Document source

The primary document source for this study was MEDLINE, the premier database of the National Library of Medicine, which includes more than eighteen million citations, representing the biomedical literature from 1949 to the present [20].

### Semantic MEDLINE

Semantic MEDLINE [9] is a multiple-step tool in development that helps users manage the results of PubMed searches. The application extracts the succinct meaning of the text it processes and displays the resulting distilled data in an interactive graph that maintains links to the original text. Semantic MEDLINE proceeds in four steps: PubMed searching, extraction of semantic predications with SemRep, automatic summarization, and visualization (Figure 1).

### SemRep

At the core of Semantic MEDLINE is SemRep [10, 11], a rule-based, symbolic natural language processing application that uses the Unified Medical Language System (UMLS) [21] to express the meaning of text in a straightforward and consistent representation, called a semantic predication. Such a representation has arguments and a predicate. The following illustrates this process:

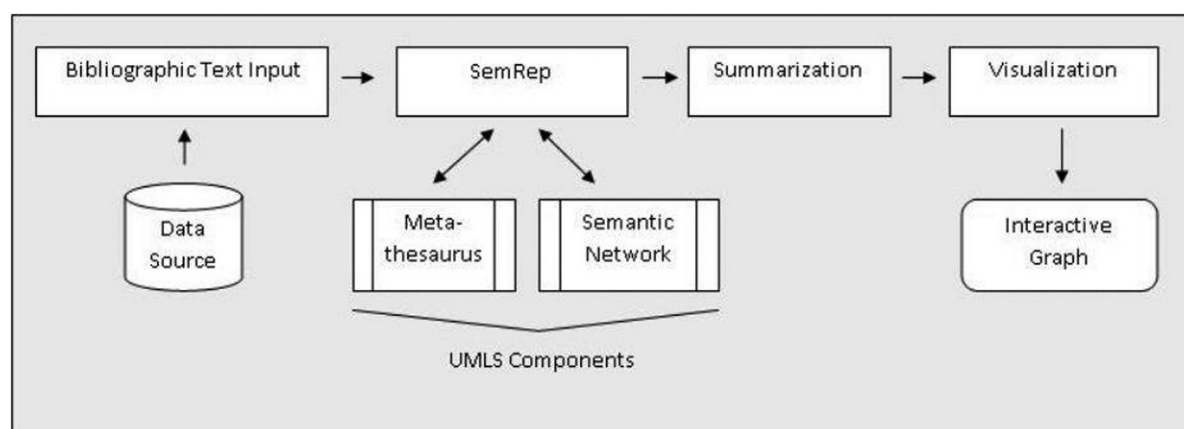
Original text: "The *IGF1R* is up-regulated in bladder cancer compared with non-malignant bladder, and might contribute to a propensity for invasion." [22]

Extracted semantic predication: *IGF1R* gene ASSOCIATED\_WITH Carcinoma of bladder

SemRep uses MetaMap [23] to map the text *IGF1R* and *bladder cancer* to the metathesaurus concepts "IGF1R gene" and "Carcinoma of bladder," which are associated with *semantic types* (or classes) "Gene or Genome" and "Neoplastic Process," respectively. These concepts function as the arguments of the predication. Based on the semantic types, SemRep then draws on the semantic network to identify the *predicate* (or relation), ASSOCIATED\_WITH, that binds these arguments. SemRep extracts semantic predications for an array of predicates, including TREATS, LOCATION\_OF, INHIBITS, INTERACTS\_

\* A public demonstration interface is at <http://skr3.nlm.nih.gov/SemMedDemo/>.

**Figure 1**  
Semantic MEDLINE



WITH, CAUSES, PREDISPOSES, and ASSOCIATED\_WITH, among others.

### Automatic summarization

In the summarization phase, a schema filters semantic predications extracted from MEDLINE citations according to a user-selected *point of view* and *topic concept* [12]. For example, if a user were interested only in information addressing treatment (i.e., the point of view) for a particular disease (i.e., the topic concept), summarization would collect the best predications that expressed this information. The summarization architecture does this by subjecting SemRep predications to four sequential phases of filtering, which select only those semantic predications pertinent to the selected point of view and topic concept:

- Relevance collects predications addressing the user-selected topic concept.
- Connectivity augments relevancy predications with others associated with the topic concept.
- Novelty eliminates predications asserting basic knowledge that users already know.
- Saliency limits final output to predications that occur most frequently.

The current online Semantic MEDLINE prototype includes schemas that summarize for treatment [12], substance interactions [24], diagnosis, and pharmacogenomics [25] points of view.

## METHODS

To explore Semantic MEDLINE's ability to assist librarians in curating secondary genetics databases, a new summarization schema was first created, targeting semantic predications that are relevant to the genetic etiology of disease. Subsequently, documents retrieved from MEDLINE were processed in the Semantic MEDLINE model enhanced with this schema. Finally, the genes identified during this

processing were evaluated by comparing them to a reference standard compiled from Genetics Home Reference and OMIM.

### A summarization schema for genetic etiology of disease

As noted earlier, a schema provides a general means of identifying SemRep predications for a particular point of view. Earlier work [26, 27] had enhanced SemRep to extract semantic predications on the genetic etiology of disease but had not provided a summarization schema. A schema for this purpose has two features: a list of allowable predicates and a list of semantic types that specify which metathesaurus concepts the listed predicates are permitted to have as arguments. The new schema was designed in such a way as to summarize SemRep data for any disease topic the user may choose, from the point of view of genetic disease etiology.

In crafting the schema, allowable semantic types were assembled into three groups: "Genetic Phenomenon," "Anatomy," and "Disease Process." The following indicates the UMLS semantic types included in each of these groups:

**Genetic Phenomenon:** Amino Acid Sequence; Enzyme; Genetic Function; Nucleic Acid, Nucleoside, or Nucleotide; Nucleotide Sequence; Amino Acid, Peptide, or Protein; Gene or Genome; and Molecular Sequence.

**Anatomy:** Anatomical Structure; Body Part, Organ, or Organ Component; Cell; Cell Component; Embryonic Structure; Fully Formed Anatomical Structure; Gene or Genome; and Tissue.

**Disease Process:** Acquired Abnormality; Anatomical Abnormality; Congenital Abnormality; Cell or Molecular Dysfunction; Disease or Syndrome; Injury or Poisoning; Mental or Behavioral Dysfunction; Neoplastic Process; Pathologic Function; Sign or Symptom; Biologic Function; Cell Function; Mental Process; Molecular Function; Natural

Phenomenon or Process; Organism Function; Organ or Tissue Function; Physiologic Function; Behavior; Mental or Behavioral Dysfunction; and Finding.

The schema for genetic etiology of disease allows the following predicates: AFFECTS, ASSOCIATED\_WITH, AUGMENTS, CAUSES, DISRUPTS, COEXISTS\_WITH, INHIBITS, PREDISPOSES, and STIMULATES. When the arguments of these predicates are limited to the semantic types noted above, the schema specifies the semantic predications permitted in summarization when generated from the point of view of the genetic etiology of disease. The following illustrates the specific semantic types (by the previously noted groups) and predicate combinations allowed by the schema:

```
{genetic phenomenon} AFFECTS {disease process}
{genetic phenomenon} AUGMENTS {disease process}
{genetic phenomenon} DISRUPTS {disease processes OR anatomy}
{genetic phenomenon} ASSOCIATED_WITH {disease process}
{genetic phenomenon} PREDISPOSES {disease process}
{genetic phenomenon} CAUSES {disease process}
{genetic phenomenon} STIMULATES {genetic phenomenon}
{genetic phenomenon} INHIBITS {genetic phenomenon}
{disease process} COEXISTS_WITH {disease process}
```

For example, this schema allows the genetic etiology predication "NAT 2 gene PREDISPOSES Carcinoma of bladder" to be included in the summary because the predicate PREDISPOSES matches, and further, the subject argument "NAT 2 gene" has the semantic type "Gene or Genome," which is included in the "genetic phenomenon" group, and the object argument has the semantic type "Neoplastic Process," which is in the "Disease Process" group. The use of three semantic groups permits predications in the summary that do not strictly assert genetic etiology but rather provide likely valuable additional information, such as "{genetic phenomenon} DISRUPTS {anatomy}" and "{disease process} COEXISTS\_WITH {disease process}." Finally, the predication "Immunotherapy TREATS Carcinoma of bladder" is not allowed, because the predicate TREATS is not in the schema.

### Acquisition of input text

To test the efficiency of the Semantic MEDLINE model (enhanced with the new schema) in identifying research literature relevant to curation of a secondary resource, the team chose bladder cancer, the sixth overall leading form of cancer in the United States [28], as a topic of study. To complete the first phase in the Semantic MEDLINE model, the project team executed the following PubMed query:

```
urinary bladder neoplasms[mh] OR "bladder cancer" OR "cancer of the bladder"
```

Limits: Publication date from 2003/01/01 to 2008/07/31, only items with abstracts, English

Five thousand six hundred six citations (titles and abstracts) were retrieved with this query and subsequently downloaded from MEDLINE.

### Document processing

All citations were processed by SemRep, and the extracted predications were then submitted to the new schema for summarization on the topic of bladder cancer according to the genetic etiology of disease point of view.

### Extraction of a list of genes from the summarized predications

A list of genes implicated in bladder cancer was extracted from the predications in the summarization schema's output, subject to the following criteria: The subject concept must have a semantic type belonging to the group "Genetic Phenomenon," and the object must be a concept referring to bladder cancer ("Carcinoma of bladder," "Bladder Neoplasm," and "Carcinoma, Transitional Cell"). These bladder cancer concepts map to the semantic type "Neoplastic Process," which is in the "Disease Process" group. For example, "FGFR3 gene" is extracted from the "FGFR3 gene ASSOCIATED\_WITH Carcinoma, Transitional Cell."

### Compilation of the reference standard from Online Mendelian Inheritance in Man and Genetics Home Reference

The reference standard for this project consisted of the genes noted as associated with bladder cancer in OMIM and Genetics Home Reference. To identify valid genes in OMIM, the team retrieved all records that were either phenotypically relevant to bladder cancer or provided clinical synopses for this disease using the following query:

```
"bladder cancer"[All Fields] OR "bladder cancers"[All Fields] OR "bladder cancer cases"[All Fields] OR "bladder cancer cell"[All Fields] OR "bladder cancer patients"[All Fields] OR "bladder carcinoma"[All Fields] OR "bladder carcinogenesis"[All Fields]
```

This query was first executed with the OMIM interface limits options manipulated to retrieve a broad range of genetic information associated with bladder cancer, varying from known genes with known chromosome loci, to hypothesized loci only, to a suspected but not ascertained genetic basis. Then, the query was issued a second time after modifying the OMIM interface limits options to retrieve only records that included a clinical synopsis. The results of these two queries were then combined, resulting in fourteen records. In Genetics Home Reference, the query "bladder" retrieved records either addressing general phenotype information (with the general label "Genetic Condition") or a gene. Of these, we identified eleven records containing information relevant to the genetic basis of bladder cancer.

The twenty-five records extracted from OMIM and Genetics Home Reference were then examined for specific genes. Records were limited to those based on source literature published within the study's time-frame (January 2003 through July 2008). Genetics Home Reference records noted ten genes with disease implications, while OMIM noted seven. Four genes were noted by both databases as relevant to bladder cancer. Genes noted in each record were classified as having a *confirmed* or *possible* involvement in bladder cancer. Genes noted in the main phenotype records of each database as implicated in bladder cancer were classified as having a *confirmed* involvement. To illustrate, the FGFR3 gene received a *confirmed* classification, due to its combination with the phrase "implicated in bladder carcinogenesis" in OMIM record #109800 for bladder cancer [29] and for its presence in the Genetics Home Reference bladder cancer condition record indicating that it is "associated with bladder cancer" [30]. Genes noted in other records in certain explicit contexts (adjacent to survival rates, for example) received a *possible* classification. For example, Genetics Home Reference notes an "amplification" of the *possible*-classified ERBB3 gene "and/or overexpression of [its] protein" in bladder tumors in the ERBB3 gene record [31]. Genes tied to conflicting, uncertain, or undefined wording were also classified as *possible*. For example, Genetics Home Reference notes conflicting evidence defining the ATM gene's implication in bladder cancer [32]. Therefore, it was assigned a *possible* classification. All genes from Genetics Home Reference and OMIM, regardless of classification, were included in the final reference standard as implicated in bladder cancer. Using these criteria, thirteen genes were included in the reference standard (Table 1).

## Evaluation

The second author (Fizman) manually matched the output of the genes extracted from the final summarization output against the genes in the reference standard. Based on this matching, recall, precision, and F-measure were calculated. Recall was defined as the percentage of genes in the reference standard that were found in the summarized output. Precision was measured by determining the percentage of all genes in the summarized output that was noted in the reference standard or in an Entrez Gene [33] Gene References into Function (GeneRIF) as implicated in bladder cancer development. GeneRIF annotations [34] in corresponding Entrez Gene records (for *Homo sapiens* only) were consulted for such genes that were not noted in OMIM or Genetics Home Reference. If an explicit GeneRIF annotation noted an association of the gene with bladder cancer, it was counted as a true positive in the precision computation. The F-measure, which ranges from a high of one to a low of zero, expresses a balanced average between the recall and precision scores.

**Table 1**

Gold standard genes associated with bladder cancer

Gene symbol	Source	Classification
FGFR3	Both	Confirmed
XPD	OMIM	Confirmed
RAG1	OMIM	Confirmed
TP53	Both	Confirmed
MTCYB	OMIM	Confirmed
HRAS	Both	Confirmed
NAT2	Both	OMIM Confirmed; GHR Possible
RB1	GHR	Confirmed
TSC1	GHR	Confirmed
ATM	GHR	Possible
TGFB1	GHR	Possible
MDM2	GHR	Possible
ERBB3	GHR	Possible

## RESULTS

SemRep extracted 38,498 semantic predications from the 5,606 citations retrieved from MEDLINE. The summarization phase limited these to 359 semantic predications relevant to bladder cancer (using the schema for genetic etiology). From these predications, 17 genes and proteins were extracted based on the criteria noted in "Extraction of a List of Genes from the Summarized Predications." These were normalized to the gene name in Entrez Gene and are shown in Table 2.

Table 3 shows the results of manually comparing the genes from summarization to the reference standard (OMIM and Genetics Home Reference) to compute recall and to Entrez Gene GeneRIFs in addition to the reference standard for computing precision. Of the thirteen genes in the reference standard, six were represented in the final summarization output. Out of seventeen genes in the summarization output, eleven were false positives when compared only to the reference standard, while only two were false positives when compared to the reference standard and GeneRIFs.

## DISCUSSION

The modified summarization system described in this paper and evaluated with bladder carcinoma genes obtained moderately good recall when compared to the reference standard compiled from OMIM and Genetics Home Reference. Precision increased substantially when GeneRIFs were taken into account. GeneRIF annotations are routinely added to an Entrez Gene record when the linked PubMed record is indexed, as part of an indexer's work, and can provide additional insight into a gene's involvement in a disease process.

There are two reasons for the level of current results. SemRep processing contributed to some errors, and further development to improve the accuracy of this application is part of ongoing research. In addition, genes are noted as implicated in a disease process in OMIM and Genetics Home Reference due to curation decisions that are, in part, independent of what is noted in the collective

**Table 2**  
Genes extracted by the summarization program

Summarization output
<b>TP53 gene*</b>
<b>FGFR3 gene*</b>
BIRC5 gene
Cadherins (CDh1)†
Cyclooxygenase 2 (PTGS2)†
CDKN2A gene
CDC91L1 gene
Candidate disease gene
<b>NAT2 gene*</b>
EGF gene
<b>TGFB1 protein, human (TGFB1)*†</b>
<b>MDM2 gene*</b>
<b>HRAS gene*</b>
GSTT1 gene
GSTM1 gene
Gelatinase B (MMP9)†
CD82 gene

\* Genes that appear in the reference standard associated with bladder cancer are in bold.  
† Genes normalized from proteins are presented in parentheses.

professional literature (and hence in SemRep output). GeneRIFs, on the other hand, are routinely created as part of the indexing process for all MEDLINE citations that include gene information. For example, the “CDC91L1 gene” was commonly noted as related to bladder cancer in the summarized SemRep output but was not noted in the OMIM and Genetics Home Reference records consulted in creating the reference standard, even though one of the GeneRIFs in Entrez Gene for CDC91L1 in *Homo sapiens* notes the following: “CDC91L1 (PIG-U) is a newly discovered oncogene in human bladder cancer” (PMID: 15034568, published within the time frame of this study). In an actual application, summarized output could guide curation, but it would be up to curators to decide what information would be included in their secondary databases.

The Semantic MEDLINE process—implementing SemRep, summarization, and visualization—converts large amounts of data into a concise representation of semantic predications expressing the data’s meaning, which can then be quickly reviewed and traced back to the original text. This process can potentially save time for database curators reviewing large amounts of information (although the project did not test this hypothesis).

Using the modified schema presented in this paper, the genetic summary can be displayed in Semantic MEDLINE as an interactive graph [9] (Figure 2). Arcs (the lines connecting the labeled concepts) represent relations between each argument node (the labeled concepts). The central node in the graph represents the user-determined topic of the summary (“Carcinoma of bladder”). The user may select or deselect predicates in the upper-right side panel, to focus on specific relationships in the graph. By right clicking on a given arc, the user can access the original text from which a semantic predication was extracted. As shown in Figure 2, the user may right-click the “PREDISPOSES” relationship arc between the GSTT1 gene concept node and the central concept “Carcino-

**Table 3**  
Performance measures\* for the summarization system on extracting genes related to bladder cancer from MEDLINE

Metric	Results
Precision	88%
Recall	46%
F-measure	0.61

\* The table displays the results with taking Gene References into Function (GeneRIFs) into account for assessing precision (as explained in “Methods”).

ma of bladder” to view the original text (a MEDLINE citation).

As noted in the introduction, use of this tool creates the potential for collaborative curation work between librarians and researchers. The following scenario further illustrates how this might work in practice: The board that oversees the institutional repository at a major university decides to integrate into this repository primary data from a university laboratory exploring the genetic etiology of disease. The librarian in charge of repository curation notes that an added-value resource summarizing the published findings of the laboratory’s research would assist other campus scientists in appraising the data. The librarian submits a query to Semantic MEDLINE to locate and download all relevant citations published by the laboratory’s faculty. The librarian then uses the application to sequentially summarize the MEDLINE data for each disease studied, from the point of view of genetic etiology. To review the summarized results, the librarian visualizes the data for each disease, clicking on the arcs in the graph to view citations associated with each semantic predication. Using the summarized data, the librarian creates a concise report of the findings associated with the lab’s data. The report is stored in the institutional repository with the lab’s research data, so that users can quickly determine its potential relevance in their own endeavors.

**Limitations of the study**

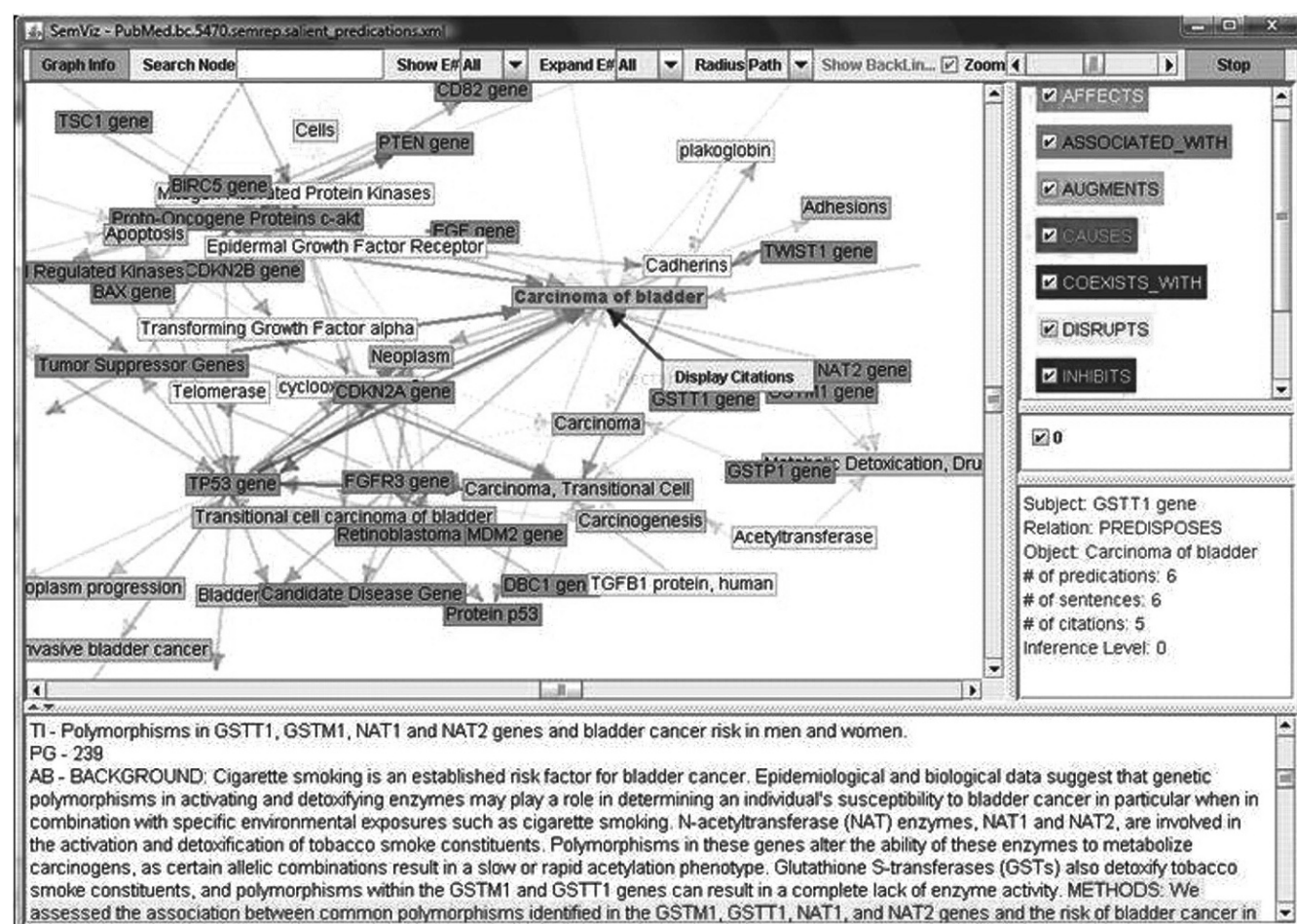
The evaluation was performed with one disease, and it is hard to predict the generalizability of performance when more diseases are taken into account. However, SemRep and the summarization system components of Semantic MEDLINE have been proved to be effective in a topic-oriented evaluation study to support evidence-based medical treatment of fifty diseases [35]. Performance will likely scale similarly to potentially support genetic database curation.

A further limitation is that the natural language processing system (SemRep) does not have access to information curators use to decide what genes are established markers for diseases. These curation policies go beyond any language processing system.

**CONCLUSIONS**

Semantic MEDLINE transforms vast amounts of bibliographic text into succinct, brief statements. To

**Figure 2**  
Visualization graph illustrating summarized semantic predications



place this in a quantitative perspective, in this study Semantic MEDLINE reduced 5,606 MEDLINE citations to 359 semantic predications. Curators could substantially reduce the amount of time needed to manually review original MEDLINE documentation by first processing it with Semantic MEDLINE and then reviewing its output.

This study explored the application of Semantic MEDLINE to a specific task, that of database curation. As noted before, this task is relevant to emerging opportunities for librarians to contribute as professional partners to parent organizations and the scientific community at large. Other work can also be aided by Semantic MEDLINE applications. For example, librarians could assist patrons in quickly assessing large amounts of bibliographic text by first processing it with Semantic MEDLINE and then instructing them on using its interactive visual display. Outcomes from separate groups of research studies, represented as bibliographic text, could be compared. These services could reaffirm the importance of university library services and strengthen the role of librarians as essential partners

in the research endeavors of their individual institutions.

Future work in schema development and domain exploration is needed to extend Semantic MEDLINE's capabilities and to measure its effectiveness. Summarization that accommodates points of view beyond those currently available will enable the system to process data for additional needs. Assessing Semantic MEDLINE's ability to assist in additional tasks such as point-of-care information delivery and patient education will give further insight to its potential uses.

## ACKNOWLEDGMENTS

The authors express their gratitude to Graciela Rosemblat for assistance with the study's evaluation, to Jeanne Marie Le Ber for editorial assistance, and to Joyce Mitchell for advice and suggestions. They also thank the National Library of Medicine for funding this project through grant number T15LM007123 and other program funding, and the Oak Ridge Institute for Science and Education for administering part of the funding.

## REFERENCES

- MacDonald S, Uribe LM. Libraries in the converging worlds of open data, e-research, and Web 2.0. Online. 2008 Mar/Apr;32(2):36–40.
- Digre KB, Lombardo NT, Frohman L. Neuro-ophthalmology Virtual Education Library (NOVEL). Neuro-Ophthalmol. 2007 Oct;31(5):175–8.
- Koopman A, Kipnis D. Feeding the fledgling repository: starting an institutional repository at an academic health sciences library. Med Ref Serv Q. 2009 Summer;28(2):111–22.
- Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nature Rev Genet. 2006 Feb 7:119–29.
- Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J. DATF: a database of Arabidopsis transcription factors. Bioinform. 2005 May 15;21(10):2568–9.
- Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J. DRTF: a database of rice transcription factors. Bioinform. 2006 May 15;22(10):1286–7.
- Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjana V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A. Development of Human Protein Reference Database as an initial platform for approaching systems biology in humans. Genome Res. 2003 Oct;13(10):2363–71.
- Caspi R, Fulcher C, Ingraham J, Keseler I, Krummenacker M, Paley S. Curator's guide for pathway/genome databases [Internet]. Menlo Park, CA: SRI International; Jul 2007 [cited 1 Jul 2010]. <<http://bioinformatics.ai.sri.com/pools/curatorsguide.pdf>>.
- Kilicoglu H, Fiszman M, Rodriguez A, Shin D, Ripple A, Rindflesch TC. Semantic MEDLINE: a web application for managing the results of PubMed Searches. Proceedings of the Third International Symposium for Semantic Mining in Biomedicine; 2008. p. 69–76.
- Rindflesch TC, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. In: Chen H, Fuller SS, Hersh W, eds. Medical informatics: knowledge management and data mining in biomedicine. Springer; 2005. p. 399–422.
- Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462–77.
- Fiszman M, Rindflesch TC, Kilicoglu H. Abstraction summarization for managing the biomedical research literature. Proceedings of the Human Language Technology/North American Chapter of the Association for Computational Linguistics Workshop on Computational Lexical Semantics; 2004. p. 76–83.
- Roberts PM. Mining literature for systems biology. Briefings Bioinform. 2006 Oct;7(4):399–406.
- National Library of Medicine. Genetics Home Reference [Internet]. Bethesda, MD: The Library; 2003 [updated 7 Apr 2004; cited 17 Dec 2009]. <<http://www.ghr.nlm.nih.gov>>.
- Mitchell JA, McCray AT. The Genetics Home Reference: a new NLM consumer health resource. AMIA Annu Symp Proc. 2003. p. 936.
- Mitchell JA, Fun J, McCray AT. Design of Genetics Home Reference: a new NLM consumer health resource. JAMIA. 2004 Dec;11(6):439–47.
- McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, and National Center for Biotechnology Information, National Library of Medicine. Online Mendelian Inheritance in Man, OMIM [Internet]. Baltimore, MD: The Library; 1995 [cited 17 Dec 2009]. <<http://www.ncbi.nlm.nih.gov/omim/>>.
- Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res. 2009 Jan;37(database issue):D793–6. Epub 2008 Oct 8.
- Letovsky SI, ed. Bioinformatics: databases and systems. Boston, MA: Kluwer Academic Publishers; 1999.
- National Library of Medicine. Data, news and update information: PubMed update [Internet]. Bethesda, MD: The Library; 2001 [updated 19 Apr 2010; cited 20 Apr 2010]. <[http://www.nlm.nih.gov/bsd/revup/revup\\_pub.html#med\\_update](http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update)>.
- Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281–91.
- Rochester MA, Patel N, Turney BW, Davies DR, Roberts IS, Crew J, Protheroe A, Macaulay VM. The type 1 insulin-like growth factor receptor is over-expressed in bladder cancer. BJU Int. 2007 Dec;100(6):1396–401.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001. p. 17–21.
- Fiszman M, Rindflesch TC, Kilicoglu H. Summarizing drug information in MEDLINE citations. AMIA Annu Symp Proc. 2006. p. 254–8.
- Ahlers CB, Fiszman M, Demner-Fushman D, Lang FM, Rindflesch TC. Extracting semantic predications from MEDLINE citations for pharmacogenomics. Pac Symp Biocomput. 2007. p. 209–20.
- Rindflesch TC, Libbus B, Hristovski D, Aronson AR, Kilicoglu H. Semantic relations asserting the etiology of genetic diseases. AMIA Annu Symp Proc. 2003. p. 554–8.
- Libbus B, Kilicoglu H, Rindflesch TC, Mork JG, Aronson AR. Using natural language processing, Locus Link, and the Gene Ontology to compare OMIM to MEDLINE. Proceedings of the Human Language Technology/North American Chapter of the Association for Computational Linguistics Workshop on Linking the Biological Literature, Ontologies and Databases: Tools for Users; 2004. p. 69–76.
- US Cancer Statistics Working Group, US Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute. United States cancer statistics: 1999–2006 incidence and mortality data [Internet]. Atlanta, GA: The Department, The Centers, and The Institute; 2009 [cited 18 Dec 2009]. <<http://www.cdc.gov/uscs/>>.
- Online Mendelian Inheritance in Man, OMIM, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, and National Center for Biotechnology Information, National Library of Medicine. #109800 Bladder cancer [Internet]. Baltimore, MD: The Library; 1995 [cited 29 Jun 2010]. <<http://www.ncbi.nlm.nih.gov/omim/109800>>.
- National Library of Medicine. Genetics Home Reference: bladder cancer [Internet]. Bethesda, MD: The Library;

- 2007 [published 13 Dec 2009; cited 18 Dec 2009]. <<http://www.ghr.nlm.nih.gov/condition=bladdercancer>>.
31. National Library of Medicine. Genetics Home Reference: ERBB3 [Internet]. Bethesda, MD: The Library; 2009 [published 13 Dec 2009; cited 18 Dec 2009]. <<http://www.ghr.nlm.nih.gov/gene=erbb3>>.
32. National Library of Medicine. Genetics Home Reference: ATM [Internet]. Bethesda, MD: The Library; 2008 [published 13 Dec 2009; cited 18 Dec 2009]. <<http://www.ghr.nlm.nih.gov/gene=atm>>.
33. National Library of Medicine. Entrez Gene [Internet]. Bethesda, MD: The Library; 2004 [cited 18 Dec 2009]. <<http://www.ncbi.nlm.nih.gov/gene/>>.
34. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM. Gene indexing: characterization and analysis of NLM's GeneRIFs. AMIA Annu Symp Proc. 2003. p. 460-4.
35. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5): 801-13.

## AUTHORS' AFFILIATIONS

**T. Elizabeth Workman, MLIS**, Graduate Student (PhD) (corresponding author), [liz.workman@utah.edu](mailto:liz.workman@utah.edu), Department of Biomedical Informatics, University of Utah, 26 S 2000 E, HSEB 5700, Salt Lake City, UT 84112; **Marcelo Fiszman, MD, PhD**, [fiszmanm@mail.nih.gov](mailto:fiszmanm@mail.nih.gov), Research Scientist, National Library of Medicine, 8600 Rockville Pike, Building 38A, Room B1N-28J, Bethesda, MD 20894; **John F Hurdle, MD, PhD**, [john.hurdle@utah.edu](mailto:john.hurdle@utah.edu), Associate Professor, Department of Biomedical Informatics, University of Utah, 26 S 2000 E, HSEB 5700, Salt Lake City, UT 84112; **Thomas C Rindflesch, PhD**, [trc@nlm.nih.gov](mailto:trc@nlm.nih.gov), Principal Investigator, Semantic Knowledge Representation Project, National Library of Medicine, 8600 Rockville Pike, Building 38A, Room 9N-913, Bethesda, MD 20894

*Received January 2010; accepted May 2010*